METHOD AND SYSTEM FOR AUTOMATED OUTLYING FEATURE AND OUTLYING FEATURE BACKGROUND DETECTION DURING PROCESSING OF DATA SCANNED FROM A MOLECULAR ARRAY

5   TECHNICAL FIELD

The present invention relates to the processing of data scanned from a molecular array and, in particular, to a method and system for automatically detecting outlying signals scanned from features and feature backgrounds based on an estimated scanned data variance calculated from the scanned data and on a maximum

10   variance threshold calculated from the scanned data and from a model variance.

BACKGROUND OF THE INVENTION

The present invention is related to processing of data scanned from molecular arrays. Molecular array technologies have gained prominence in biological

15   research and are likely to become important and widely used diagnostic tools in the healthcare industry. Currently, molecular-array techniques are most often used to determine the concentrations of particular nucleic-acid polymers in complex sample solutions. Molecular-array-based analytical techniques are not, however, restricted to analysis of nucleic acid solutions, but may be employed to analyze complex solutions

20   of any type of molecule that can be optically or radiometrically scanned and that can bind with high specificity to complementary molecules synthesized within, or bound to, discrete features on the surface of a molecular array. Because molecular arrays are widely used for analysis of nucleic acid samples, the following background information on molecular arrays is introduced in the context of analysis of nucleic

25   acid solutions following a brief background of nucleic acid chemistry.

Deoxyribonucleic acid ("DNA") and ribonucleic acid ("RNA") are linear polymers, each synthesized from four different types of subunit molecules. The subunit molecules for DNA include: (1) deoxy-adenosine, abbreviated "A," a purine nucleoside; (2) deoxy-thymidine, abbreviated "T," a pyrimidine nucleoside; (3)

30   deoxy-cytosine, abbreviated "C," a pyrimidine nucleoside; and (4) deoxy-guanosine, abbreviated "G," a purine nucleoside. The subunit molecules for RNA include: (1)

adenosine, abbreviated "A," a purine nucleoside; (2) uracil, abbreviated "U," a pyrimidine nucleoside; (3) cytosine, abbreviated "C," a pyrimidine nucleoside; and (4) guanosine, abbreviated "G," a purine nucleoside. Figure 1 illustrates a short DNA polymer 100, called an oligomer, composed of the following subunits: (1) deoxy-

5 adenosine 102; (2) deoxy-thymidine 104; (3) deoxy-cytosine 106; and (4) deoxy-guanosine 108. When phosphorylated, subunits of DNA and RNA molecules are called "nucleotides" and are linked together through phosphodiester bonds 110-115 to form DNA and RNA polymers. A linear DNA molecule, such as the oligomer shown in Figure 1, has a 5' end 118 and a 3' end 120. A DNA polymer can be chemically

10 characterized by writing, in sequence from the 5' end to the 3' end, the single letter abbreviations for the nucleotide subunits that together compose the DNA polymer. For example, the oligomer 100 shown in Figure 1 can be chemically represented as "ATCG." A DNA nucleotide comprises a purine or pyrimidine base (e.g. adenine 122 of the deoxy-adenylate nucleotide 102), a deoxy-ribose sugar (e.g.

15 deoxy-ribose 124 of the deoxy-adenylate nucleotide 102), and a phosphate group (e.g. phosphate 126) that links one nucleotide to another nucleotide in the DNA polymer. In RNA polymers, the nucleotides contain ribose sugars rather than deoxy-ribose sugars. In ribose, a hydroxyl group takes the place of the 2' hydrogen 128 in a DNA nucleotide. RNA polymers contain uridine nucleosides rather than the deoxy-

20 thymidine nucleosides contained in DNA. The pyrimidine base uracil lacks a methyl group (130 in Figure 1) contained in the pyrimidine base thymine of deoxy-thymidine.

The DNA polymers that contain the organization information for living organisms occur in the nuclei of cells in pairs, forming double-stranded DNA

25 helixes. One polymer of the pair is laid out in a 5' to 3' direction, and the other polymer of the pair is laid out in a 3' to 5' direction. The two DNA polymers in a double-stranded DNA helix are therefore described as being anti-parallel. The two DNA polymers, or strands, within a double-stranded DNA helix are bound to each other through attractive forces including hydrophobic interactions between stacked

30 purine and pyrimidine bases and hydrogen bonding between purine and pyrimidine bases, the attractive forces emphasized by conformational constraints of DNA

polymers. Because of a number of chemical and topographic constraints, double-stranded DNA helices are most stable when deoxy-adenylate subunits of one strand hydrogen bond to deoxy-thymidylate subunits of the other strand, and deoxy-guanylate subunits of one strand hydrogen bond to corresponding deoxy-cytidilate subunits of the other strand.

Figures 2A-B illustrate the hydrogen bonding between the purine and pyrimidine bases of two anti-parallel DNA strands. Figure 2A shows hydrogen bonding between adenine and thymine bases of corresponding adenosine and thymidine subunits, and Figure 2B shows hydrogen bonding between guanine and cytosine bases of corresponding guanosine and cytosine subunits. Note that there are two hydrogen bonds 202 and 203 in the adenine/thymine base pair, and three hydrogen bonds 204-206 in the guanosine/cytosine base pair, as a result of which GC base pairs contribute greater thermodynamic stability to DNA duplexes than AT base pairs. AT and GC base pairs, illustrated in Figures 2A-B, are known as Watson-Crick ("WC") base pairs.

Two DNA strands linked together by hydrogen bonds forms the familiar helix structure of a double-stranded DNA helix. Figure 3 illustrates a short section of a DNA double helix 300 comprising a first strand 302 and a second, anti-parallel strand 304. The ribbon-like strands in Figure 3 represent the deoxyribose and phosphate backbones of the two anti-parallel strands, with hydrogen-bonding purine and pyrimidine base pairs, such as base pair 306, interconnecting the two strands. Deoxy-guanylate subunits of one strand are generally paired with deoxy-cytidilate subunits from the other strand, and deoxy-thymidilate subunits in one strand are generally paired with deoxy-adenylate subunits from the other strand. However, non-WC base pairings may occur within double-stranded DNA. Generally, purine/pyrimidine non-WC base pairings contribute little to the thermodynamic stability of a DNA duplex, but generally do not destabilize a duplex otherwise stabilized by WC base pairs. However, purine/purine base pairs may destabilize DNA duplexes.

Double-stranded DNA may be denatured, or converted into single stranded DNA, by changing the ionic strength of the solution containing the double-

stranded DNA or by raising the temperature of the solution. Single-stranded DNA polymers may be renatured, or converted back into DNA duplexes, by reversing the denaturing conditions, for example by lowering the temperature of the solution containing complementary single-stranded DNA polymers. During renaturing or

5 hybridization, complementary bases of anti-parallel DNA strands form WC base pairs in a cooperative fashion, leading to regions of DNA duplex. Strictly A-T and G-C complementarity between anti-parallel polymers leads to the greatest thermodynamic stability, but partial complementarity including non-WC base pairing may also occur to produce relatively stable associations between partially-complementary polymers.

10 In general, the longer the regions of consecutive WC base pairing between two nucleic acid polymers, the greater the stability of hybridization between the two polymers under renaturing conditions.

The ability to denature and renature double-stranded DNA has led to development of many extremely powerful and discriminating assay technologies for

15 identifying the presence of DNA and RNA polymers having particular base sequences or containing particular base subsequences within complex mixtures of different nucleic acid polymers, other biopolymers, and inorganic and organic chemical compounds. These methodologies include molecular-array-based hybridization assays. Figures 4-7 illustrate the principle of molecular-array-based hybridization

20 assays. A molecular array (402 in Figure 4) comprises a substrate upon which a regular pattern of features are prepared by various different types of manufacturing processes. The molecular array 402 in Figure 4, and in subsequent Figures 5-7, has a grid-like two-dimensional array of square features, such as feature 404 shown in the upper left-hand corner of the molecular array. Each feature of the molecular array

25 contains a large number of identical oligonucleotides covalently bound to the surface of the feature. In general, chemically distinct oligonucleotides are bound to the different features of a molecular array, so that each feature corresponds to a particular nucleotide sequence. In Figures 4-6, the principle of molecular-array-based hybridization assays is illustrated with respect to the single feature 404 to which a

30 number of identical oligonucleotides 405-409 are bound. In practice, each feature of

the molecular array contains an enormous number of oligonucleotide molecules, but, for the sake of clarity, Figures 4-6 only show a small number.

Once a molecular array has been prepared, the molecular array may be exposed to a sample solution of DNA molecules that includes DNA molecules (410-

5 413 in Figure 4) labeled with fluorophores, chemoluminescent compounds, or radioactive atoms 415-418. A labeled DNA molecule that contains a nucleotide sequence complementary to the base sequence of an oligonucleotide bound to the molecular array may hybridize through base pairing interactions to the oligonucleotide. Figure 5 shows a number of labeled DNA molecules 502-504

10 hybridized to oligonucleotides 505-507 bound to the surface of the molecular array 402. DNA molecules that do not contains nucleotide sequences complementary to any of the oligonucleotides bound to the molecular array do not hybridize stably to oligonucleotides bound to the molecular array and generally remain in solution, such as labeled DNA molecules 508 and 509. The sample solution is then rinsed from the

15 surface of the molecular array, washing away any unbound labeled DNA molecules. Finally, as shown in Figure 6, the bound labeled DNA molecules are detected via optical or radiometric scanning. Optical scanning involves exciting labels of bound labeled DNA molecules with electromagnetic radiation of appropriate frequency and detecting fluorescent emissions from the labels, or detecting light emitted from

20 chemoluminescent labels. When radioisotope labels are employed, radiometric scanning can be used to detect radiation emitted from labeled DNA molecules hybridized to oligonucleotides bound to the surface of the molecular array. Additional types of signals are also possible, including electrical signals generated by electrical properties of bound target molecules, magnetic properties of bound target

25 molecules, and other such physical properties of bound target molecules that that can produce a detectable signal. Optical, radiometric, or other types of scanning produce an analog or digital representation of the molecular array as shown in Figure 7, with features to which labeled DNA molecules are hybridized similar to 706 optically or digitally differentiated from those features to which no labeled DNA molecules are

30 bound. In other words, the analog or digital representation of a scanned molecular array displays positive signals for features to which labeled DNA molecules are

hybridized and displays negative features to which no, or an undetectably small number of, labeled DNA molecules are bound. Features displaying positive signals in the analog or digital representation indicate the presence of DNA molecules with complementary nucleotide sequences in the original sample solution. Moreover, the

5   signal intensity produced by a feature is generally related to the amount of labeled DNA bound to the feature, in turn related to the concentration, in the sample to which the molecular array was exposed, of labeled DNA complementary to the oligonucleotide within the feature.

Molecular-array-based hybridization techniques allow extremely

10   complex solutions of DNA molecules to be analyzed in a single experiment. Molecular arrays may contain hundreds, thousands, or tens of thousands or different oligonucleotides, allowing for the detection of hundreds, thousands, or tens of thousands of different DNA polymers containing complementary nucleotide sub-sequences in the complex DNA solutions to which the molecular array is exposed. In

15   order to perform different sets of hybridization analyses, molecular arrays containing different sets of bound oligonucleotides are manufactured by any of a number of complex manufacturing techniques. These techniques generally involve synthesizing the oligonucleotides within corresponding features of the molecular array through complex iterative synthetic steps.

20   As pointed out above, molecular-array-based assays can involve other types of biopolymers, synthetic polymers, and other types of chemical entities. For example, one might attach protein antibodies to features of the molecular array that would bind to soluble labeled antigens in a sample solution. Many other types of chemical assays may be facilitated by molecular array technologies. For example,

25   polysaccharides, glycoproteins, synthetic copolymers, including block coploymers, biopolymer-like polymers with synthetic or derivitized monomers or monomer linkages, block copolymers, and many other types of chemical entities may serve as probe and target molecules for molecular-array-based analysis. A fundamental principle upon which molecular arrays are based is that of specific recognition, by

30   probe molecules affixed to the molecular array, of target molecules, whether by sequence-mediated binding affinities, binding affinities based on conformational or

topological properties of probe and target molecules, or binding affinities based on spatial distribution of electrical charge on the surfaces of target and probe molecules.

DNA, and other biological polymers, may be labeled with different chemical chromophores, radioactive nuclides, or other signal-generating entities, and may be optically scanned at different wavelengths of light, radiometrically scanned for different types of radioactive emission within different energy ranges, or scanned by other techniques appropriate to detect signals produced by other signal-generating entities. In the case of optical scanning, each different wavelength at which a molecular array is scanned produces a different signal. Thus, in optical scanning, it is common to describe the signal produced by scanning in terms of the color of the wavelength of light employed for the scan. For example, a red signal is produced by scanning a molecular array with light having a wavelength corresponding to that of visible red light.

Scanning of a feature by an optical scanning device or radiometric scanning device generally produces a scanned image comprising a rectilinear grid of pixels, with each pixel having a corresponding signal intensity. Figure 8A shows a portion of a scanned image of a molecular array that includes a pixel-based image of a disk-shaped feature of a molecular array. In Figure 8A, the feature corresponds to a disk-shaped region 802 of pixels having relatively high signal intensities. Surrounding the feature 802 is a ring-like region 804 of pixels with relatively low measured intensities. The portion of the scanned image shown in Figure 8A is thus conceptually equivalent to a digital, black-and-white photograph of the feature taken with light within a narrow range of wavelengths. Generally, the location of the disk-shaped region 802 corresponding to a feature is determined by various scanned image-to-scanned-molecular-array alignment techniques and procedures.

It is desirable for the signal intensities, or counts, of pixels within the area of a pixel-based scanned image corresponding to a feature to be relatively uniform. Similarly, it is also desirable for the signal intensities within background regions surrounding features to be relatively uniform. Non-uniform signal intensity distributions generally indicate the occurrence of one or more error or noise conditions that may prevent meaningful data from being collected from the feature.

Figures 8B-D illustrate various non-uniform signal intensity distributions within a scanned image of a molecular array feature. In Figure 8B, for example, relatively large signal intensities are seen in regions 806 and 808 at the upper right, and lower left, of the scanned image as well as within the disk-shaped

5   area 810 corresponding to a feature. Such non-uniform distribution of signal intensities may indicate defects in the preparation of the molecular array, including defects in the synthesis of probe molecules bound to the molecular array, contamination of the surface of the molecular array with a chromophore that responds to impinging light in a similar fashion to the response by the chromophore with which

10   target molecules are labeled, flaws in the scanning device, or other such defects. In Figure 8C, the signal intensities within the feature 812 are relatively uniform, with the exception of a number of extremely high, outlying signal intensities in individual pixels, such as pixels 814, 816, and 818. Such outlying pixel intensities may represent scanner measurement errors or defects in digital processing and digital

15   representation of the scanned data. In Figure 8D, a relatively large area 820 within a feature 822 has produced no signal, and therefore represents a significant spatial non-uniformity of pixel intensities. A condition such as that shown in Figure 8D may arise when probe molecules are not uniformly bound to the surface of the molecular array within a feature, because of overlying contamination that masks the signal, or

20   for other reasons. In the situations illustrated in Figures 8B-D, the sum of the pixel intensities within the disk-shaped region of the optical image corresponding to a feature may produce a total signal intensity, or count, for the feature that does not reflect the theoretical count that would be produced by scanning the feature were the one or more error conditions or noise conditions not present. Such scanned features

25   suffering from non-uniform pixel intensities need to be recognized during processing of data scanned from a molecular array and flagged as outlier features, to prevent reporting of flawed and erroneous experimental results.

Currently, outlier features, or feature backgrounds, are commonly identified by using negative control features manufactured into molecular arrays and

30   by manual inspection of scanned images. However, control-feature-based outlier detection may be insensitive to various types of non-uniformities and significantly

adds to the cost of molecular array manufacture and molecular array scanning and data processing. Manual outlier detection suffers from the inaccuracies and deficiencies well-known to occur in most human-dependent tasks, and is also quite slow and economically inefficient. Thus, designers, manufacturers, and users of

5   molecular arrays have recognized the need for a more accurate, automated technique for recognizing outlier features and outlier feature backgrounds in scanned images of molecular arrays.

SUMMARY OF THE INVENTION

10   The present invention is directed towards a method and system for identifying outlier features and outlier feature backgrounds in scanned images of molecular arrays. The method and system of the present invention employ pixel-based, signal-intensity data contained within areas of a scanned image of a molecular array corresponding to features and feature backgrounds in order to determine

15   whether or not the features or feature backgrounds have non-uniform signal intensities and are thus outlier features and outlier feature backgrounds. A calculated, estimated variance for the signal intensities within a feature or feature background is compared to a maximum allowable variance calculated for the feature or feature background based on a signal intensity variance model. When the experimental

20   variance is less than or equal to the maximum allowable variance, the feature or feature background is considered to have acceptable signal-intensity uniformity. Otherwise, the feature or feature background is flagged as an outlier feature or outlier feature background.

25   BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 shows a linear DNA polymer.

Figures 2A-B   illustrate   the   hydrogen   bonding   between purine/pyrimidine bases of two anti-parallel DNA strands.

Figure 3 illustrates a short section of a DNA double helix.

30   Figures 4-7   illustrate   the   principle   of   molecular-array-based hybridization assays.

Figure 8A shows a pixel-based result from scanning a disk-shaped feature of a molecular array.

Figures 8B-D illustrate various non-uniform signal intensity distributions within a scanned optical image of a molecular array feature.

5   Figure 9A shows a generalized normal distribution.

Figure 9B shows a binomial distribution.

Figure 9C shows a representative χ^2 distribution.

10   DETAILED DESCRIPTION OF THE INVENTION

The present invention is directed to identifying outlier features and outlier feature backgrounds within scanned images of molecular arrays. The variance of signal intensities within a feature or feature background is compared to a maximum allowable variance calculated based on a variance model in order to

15   determine whether or not the region of a scanned image of a molecular array corresponding to a feature or feature background contains adequately uniform pixel-based signal intensities within. In the following, a description of the variance model and the fundamental statistical concepts and distributions on which it is based is provided with reference to Figures 9A-C and a number of mathematical formulas.

20   Following this discussion, a C++-like pseudocode implementation of automated outlier detection functionality that may be embedded within a molecular-array data processing system is provided as a described embodiment of the present invention.

Data processing techniques employed in outlier detection involve application of various statistical measurements on the per-pixel counts, or pixel-based

25   signal intensities measured for a particular feature or feature background and included in a digital representation of the scanned image of the molecular array. A molecular array scanner produces a raw digital representation including a count, or signal intensity, for each pixel within the digital representation. As a first step in processing the raw data, net signals "S_{net}" are calculated from measured signals "$S_{measured}$" via a

30   subtractive process:

$$S_{net} = S_{measured} - S_{offset}$$

For each measured per-pixel count, or pixel-based signal intensity, the net signal is obtained by subtracting a signal offset "S_{offset}" from the measured signal "$S_{measured}$."

5 The signal offset may be automatically provided by the scanner device or may be empirically determined by identifying a minimal signal in the digital representation of the molecular array produced by scanning the molecular array and processing the scanned data. An estimate of the variance of the per-pixel counts within the area of a digital representation of a molecular array corresponding to a feature or feature

10 background is obtained as follows:

$$S^2{}_{s_{net}} = \frac{1}{n-1} \sum_{i=1}^{n} \left(s_{net} - \bar{s}_{net} \right)^2$$

$$where \ \bar{s}_{net} = \frac{1}{n} \sum_{i=1}^{n} s_{net},$$

S = standard deviation, and

n = the number of pixels within the feature or feature

15 background

Thus, the variance of pixel counts or pixel-based signal intensities within a feature or feature background can be straightforwardly calculated from the net signals obtained from the digital representation of the scanned image of a molecular array.

20 In order to determine whether the pixel counts or pixel-based signal intensities within a feature or feature background are sufficiently uniform, the calculated variance "$S^2{}_{s_{net}}$" needs to be compared to a threshold value to determine whether or not the calculated variance "$S^2{}_{s_{net}}$" falls below the threshold value and therefore is acceptable. While current methods employ values measured from

25 negative control features included within a molecular array, or depend on manual inspection of pixel count distributions, the present invention employs a calculated variance model to obtain the threshold value. In one embodiment of the present

invention, the calculated variance model "$\hat{\sigma}^2$" is a linear combination of three different, independent model variances:

$$\hat{\sigma}^2 = \hat{\sigma}^2_{\text{labeling and feature synthesis}} + \hat{\sigma}^2_{\text{counting}} + \hat{\sigma}^2_{\text{noise}}$$

5

The model variance "$\hat{\sigma}^2_{\text{labeling and feature synthesis}}$" is the variance expected for non-uniformities associated with target-molecule labeling, feature synthesis, and other solution and surface and chemistry effects. The model variance "$\hat{\sigma}^2_{\text{counting}}$" is the variance expected in scanning measurement, or counting, error. The model variance

10   "$\hat{\sigma}^2_{\text{noise}}$" is the expected variance due to electronic noise in the scanner, background-level signal noise produced by the glass substrate of the molecular array, and other such noise.

In one embodiment of the present invention, the non-uniformity associated with labeling and feature synthesis is considered to be normally

15   distributed. Figure 9A illustrates a generalized normal distribution, described by the following expression:

$$f(y) = \frac{e^{-(y-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\Pi}}$$

20

where y = measured quantity,

μ = mean,

σ = standard deviation

25   In the described embodiment of the present invention, the non-uniformity associated with scanner measurement error is considered to be distributed according to a Poisson distribution. Figure 9B illustrates a binomial distribution, described by the following expression:

$$p(y) = \binom{n}{y} p^y q^{n-y}$$

where p(y) = probability of y positive outcomes,

p = probability of a positive outcome,

q = probability of a negative outcome, and

n = counts, time intervals, etc.

A Poisson distribution is the limit of the binomial distribution as n approaches infinity. The Poisson distribution is expressed as follows:

$$p(y) = \frac{\lambda^y}{y!} e^{-\lambda}$$

where $\lambda = \bar{y}$

The non-uniformity associated with electronic scanner noise and glass-substrate-background-level noise is considered to be a constant, in the described embodiment.

In the described embodiment, the model variance "$\hat{\sigma}^2$" is alternatively expressed as:

$$\hat{\sigma}^2 = A\bar{s}^2_{net} + B\bar{s}_{net} + C$$

where $A\bar{s}^2_{net} = \hat{\sigma}^2_{\text{labeling and feature synthesis}}$,

$B\bar{s}_{net} = \hat{\sigma}^2_{\text{counting}}$,

$C = \hat{\sigma}^2_{\text{noise}}$

The constant "A" can be estimated as the square of the coefficient of variation, "$\dfrac{\sigma^2}{\bar{s}^2_{net}}$", which can be estimated based on analyzing large numbers of similar molecular arrays and computing the coefficient of variation in the analysis. In the case of in-situ arrays, a value of A = 0.01 provides a good estimate for the square of the coefficient of variation due to labeling and feature synthesis non-uniformities.

For the scanner Poissonian noise, the signal to noise ratio is estimated, in the described embodiment, based on the number of molecules of chromophores and the number of photons produced by each molecule, as follows:

$$S/N = \frac{\sqrt{m}\sqrt{p}}{\sqrt{p+1}}$$

where m = number of chromophore molecules, and

p = number of photons/chromophore molecule

Therefore, when the number of photons emitted per chromophore is large, the signal to noise expression is provided below:

$$S/N \approx \sqrt{m}$$

In the described embodiment, the scanner measures a signal of approximately 3.2 counts per chromophore molecule in a 10 micron by 10 micron pixel. Therefore, the number of chromophore molecules per pixel can be estimated by the mean counts per pixel $\bar{s}_{net}$ as follows:

$$m = \frac{\bar{s}_{net}}{3.2}$$

Thus,

$$S/N = \sqrt{\frac{\bar{s}_{net}}{3.2}}$$

$$\sigma^2_{counting} = \left(\frac{1}{S/N}\right)^2 = 3.2\,\bar{s}_{net}$$

$$B = 3.2$$

In the described embodiment, the constant "C" is found, through scanning experiments, to have a value of 144. The estimated values of constants "A," "B," and "C" obviously vary with varying experimental conditions, target and probe biopolymers, molecular array substrates, chromophores, and scanning and data

5   reduction equipment.

Using the above-described variance model, a threshold value, or $\hat{\sigma}^2_{max}$, can be estimated using an assumption that the following expression is distributed according to a χ^2 distribution with n-1 degrees of freedom, where n is the number of feature or feature background pixels:

10

$$\frac{(n-1)\hat{\sigma}^2}{\sigma^2}$$

where σ^2 is the true feature or feature background variance under the assumption
that the model is valid, and the feature or feature background is not an outlier

15   A representative χ^2 distribution is shown in Figure 9C, where the χ^2 distribution is expressed as follows:

$$f(y) = \begin{cases} \dfrac{y^{(v/2)-1}e^{-y/2}}{2^{v/2}\Gamma(v/2)} & ,y \geq 0, v \geq 0 \\ 0 \end{cases}$$

$$\text{where } \Gamma(v/2) = \int_0^\infty y^{(v/2)-1}e^{-y}dy$$

$$v = \text{number of degrees of freedom}$$

20

The threshold value is determined by selecting a lower bound "χ_x" (902 in Figure 9C) such that the probability that the χ^2-distributed expression $(n-1)\dfrac{\hat{\sigma}^2}{\sigma^2}$ is greater than $1-\dfrac{\alpha}{2}$, where the probability $1-\dfrac{\alpha}{2}$ is the areas under the distribution

curve 904 to the right of the lower bound "χ_x" 902, according to the following expression:

$$p\left[\chi_x{}^2 \le \frac{(n-1)\hat{\sigma}^2}{\sigma^2}\right] = 1 - \frac{\alpha}{2}$$

5

By rearranging the above expression, an equivalent expression is obtained:

$$p\left[\sigma^2 \le \frac{(n-1)\hat{\sigma}^2}{\chi_x{}^2}\right] = 1 - \frac{\alpha}{2}$$

10   Thus:

$$\sigma_{max}{}^2 = \frac{(n-1)\hat{\sigma}^2}{\chi_x{}^2}$$

15   It should be noted that, although the above described variance model has been found to provide an effective basis for outlier detection, many other type of variance models are possible. Additional terms can be included, to account for other types of variances, terms may be modified, to more precisely describe the variances, and terms may be deleted from the above expression for the model variance. The

20   techniques of the present invention may use any of the many possible model variances for outlier detection.

          A C++-like pseudocode implementation showing an embodiment of the present invention is provided below. Note that the pseudocode implementation is not intended to describe a complete data processing program for molecular array data, but

25   only to provide sufficient detail to illustrate one possible embodiment the above-described outlier identification methodology as the embodiment might occur within a molecular array data processing program, or in molecular array scanning and data

processing equipment. The molecular array data processing program including the techniques of the present invention analyzes data scanned from a molecular array to produce experimental or diagnostic results which are stored in a computer-readable medium, transferred to an intercommunicating entity via electronic signals, printed in

5　a human-readable format, or otherwise made available for further use.

First, the pseudocode implementation includes several constants and enumerations:

```
1   const int numColors = 2;
2   enum color {RED, GREEN};

3   const int numAreas = 2;
4   enum area {FEATURE, BACKGROUND};
```

The enumeration "color" contains a value for each different type signal, or scanning wavelength, and the number of different types of signals is provided by the constant "numColors." Similarly, areas that can be counted and analyzed statistically include features and feature backgrounds, described in the enumeration "area." The constant "numAreas" describes the number of areas in the enumeration "area." Thus, the pseudocode implementation includes analysis of both red and green signals for features as well as for feature backgrounds.

Next, the pseudocode implementation includes the class "scannedData," provided below:

```
1   class scannedData
2   {
3       private:
4           int* data;
5           int rowSize;
6           int* colSize;
7           int total;
8           bool outlying;
9
10      public:
11
12          int getRowSize() {return rowSize;};
13          int getColSize(int row) {return *(colSize + row);};
```

```
14          int getPixelCount(int row, int col);
15          int getTotal();
16          void setTotal(int t) {total = t;};
17          bool getOutlying() {return outlying;};
18          void setOutlying(bool y) {outlying = y;};
19          scannedData(int* data);
20   };
```

An instance of the class "scannedData" describes the pixel-based signal intensities, or counts, for a particular background or feature area of a scanned molecular array. The pixels are assumed to be rectilinearly oriented, with the shape of the area having a major horizontal axis, or row, that intersects with all columns of pixels within the area. Thus, the pseudocode implementation can model square features, disk-shaped features, elliptically shaped features, and other similar symmetrical closed forms. The class "scannedData" contains the following data members: (1) "data," a pointer to the pixel counts; (2) "rowSize," the size, in columns, of the major horizontal axis, or major row; (3) "colSize," a pointer to the sizes of columns that include each pixel of the major row; (4) "total," a total number of counts for the area of the scanned image; and (5) "outlying," a Boolean value indicating whether or not the distribution of counts within the area is non-uniform. The class "scannedData" includes various member functions for setting and retrieving the values of the above-described data members, a member function "getPixelCount" that returns the per-pixel count measured by a scanning device by the pixel with row and column coordinates supplied as arguments, and a constructor "scannedData" that takes raw data as input. An implementation for the member function "getPixelCount" and the constructor are not provided, as the implementations are quite dependent on the format of the raw data and implementation of other portions of the data processing package, and are outside the scope of the present invention.

Next, the pseudocode implementation includes a declaration of the class "feature," provided below:

```
1   class feature
2   {
3        private:
4             int x_coordinate;
```

```
5        int y_coordinate;
6        scannedData *features;
7
8   public:
9        bool outlier(area a, color c)
10           {return features[a * numAreas + c].getOutlying();};
11       int getCount(area a, color c)
12           {return features[a * numAreas + c].getTotal();};
13       feature(scannedData* data, int* offsets,
14           float* A, float* B, float* C, float chiSquaredXPoint,
15           int x, int y);
16       virtual ~feature();
17  };
```

An instance of the class "feature" describes a feature of the molecular array, and includes a pointer to an array of instances of the class "scannedData," described above, for the areas corresponding to the feature and to the background feature scanned at red and green visible wavelengths. The class feature includes the following data members: (1) "x_coordinate," the x coordinate of the feature in a rectilinear grid of features that comprises the molecular array; (2) the y coordinate of the feature; and (3) "features," a pointer to an array of instances of the class "scannedData." The class feature includes the following member functions: (1) "outlier," declared and implemented on lines 9 and 10, above, which returns a Boolean value indicating whether or not the area of the feature corresponding to argument "a" is an outlier with respect to the signal provided by argument "c;" (2) "getCount," declared and implemented above on lines 11-12, which returns the total net signal for either the background of the feature or the feature and scanned at a particular wavelength; and (3) "feature," a constructor for the feature.

The constructor for the class "feature" contains the code relevant to one embodiment of the present invention. An implementation for the constructor "feature" is provided below:

```
1   feature::feature(scannedData* data, int* offsets,
2        float* A, float* B, float* C, float chiSquaredXPoint,
3        int x, int y)
4   {
5        int total;
6        int total2;
7        int num;
```

```
8       int count;
9       double s_net;
10      double s_net2;
11      double s2_model;
12      double s2_max;
13      double s2;
14
15      features = data;
16      x_corrdinate = x;
17      y_coordinate = y;
18
19      for (int i = 0; i < numColors; i++)
20      {
21          for (int j = 0; j <= numAreas; j++)
22          {
23              total = 0;
24              total2 = 0;
24              num = 0;
25
26              for (int k = 0; k < data->getRowSize(); k++)
27              {
28                  for (int l = 0; l < data->getColSize(k); l++)
29                  {
30                      count = data->getPixelCount(k, l) - *offsets;
31                      total2 += count * count;
32                      total += count;
33                      num++;
34                  }
35              }
36              s_net = total / num;
37              s_net2 = s_net * s_net;
38              s2_model = (s_net2 * (*A)) + (s_net * (*B)) + *C;
39              s2_max = s2_model * (num - 1) / chiSquaredXPoint;
40              s2 = total2/num - s_net2;
41              if (s2 <= s2_max) data->setOutlying(false);
42              else data->setOutlying(true);
43              data->setTotal(total);
44              data++;
45              offsets++;
46              A++;
47              B++;
48              C++;
49          }
50      }
51  }
```

The constructor "feature" takes the following arguments: (1) "data," a pointer to an array of instances of the class "scannedData;" (2) "offsets," a pointer to an array of offsets, corresponding to the term "s_{offset}" in the above-described expression for the

net signal "s_{net}"; (3) "A," "B," and "C," pointers to arrays of constants for each type of scanned in area, e.g., feature or feature background scanned in red or green light, where the constants in the arrays correspond to the constants "A," "B," and "C," in the above-described expression for the model variance "$\hat{\sigma}^2$;" (4)

5   "chiSquaredXPoint," the threshold variance value "χ^2_x," described above; and (5) "x" and "y," the x and y coordinates for the feature. On lines 5-13, a number of local variables are declared. These local variables include: (1) "total," pixel counts obtained from an area associated with a feature during a particular scan; (2) "total2," the square of the total pixel counts; (3) "num," the number of pixels in the area; (4)

10   "count," a particular net count for a pixel "s_{net};" (5) "s_net," the average value of the net signals from an area; (7) "s_net2," the square of the average net signals from an area; (8) "s2_model," the calculated model variance for an area feature under a particular scan; (8) "s2_max," the threshold value "σ^2_{max}," described above; and (9) "s2," the estimated variance for the pixel intensities within the area. On lines 15-17,

15   member data for the class feature are initialized based to the values of supplied arguments. In the nested *for*-loops of lines 19-50, each of the instances of the class "scannedData" describing scans of areas associated with the feature are processed according to the above-described technique for obtaining net signals and determining whether or not the uniformity of the signal intensities within an area are acceptable.

20   Thus, the code of lines 22-48 is executed for each scan of each areas associated with the feature. In the case of the described embodiment, instances of the class "scannedData" represent red and green scans of the feature background and the feature. In the *for*-loop of lines 26-35, the square of the total net signals, the total net signals, and the number of pixels in an area are calculated for the area. On line 36,

25   the value $\bar{s}_{net}$ is calculated. On line 37, the value $\bar{s}_{net}^{\ 2}$ is calculated. On line 38, the value $\hat{\sigma}^2$ is calculated. On line 39, the value σ^2_{max} is calculated. On line 40, the estimated variance for the pixel counts within the area is calculated. On lines 41-42, the member data "outlier" for the instance of the class "scannedData" is set to "false" if the estimated variance is less than or equal to the threshold variance σ^2_{max}, and is

30   set to "true" otherwise. On line 43, the member data "total" is set to the total net

signal count for the area. Finally, on lines 44-48, array pointers are incremented for the next iteration of the nested *for*-loops.

Although the present invention has been described in terms of a particular embodiment, it is not intended that the invention be limited to this

5 embodiment. Modifications within the spirit of the invention will be apparent to those skilled in the art. For example, an almost limitless number of different implementations of the outlier detection method of the present invention can be written in any of many different programming languages, embodied in firmware, embodied in hardware circuitry, or embodied in a combination of one or more of

10 firmware, hardware, or software, for inclusion in molecular array data processing equipment employing a computational processing engine to execute software or firmware instructions encoding techniques of the present invention or including logic circuits that embody both a processing engine and instructions. Various different variance models can be employed, including models with additional model variance

15 terms corresponding to observed errors, defects, and noises different from, in addition to, or in place of those used in the described embodiment. Use of statistical variance modeling for generating variance thresholds for outlier detection can be applied to many different types of molecular arrays, and to many other molecular-array-like scientific and diagnostic devices. In the described embodiment, the

20 techniques of the present invention are employed to detect outlier features and features backgrounds, but the same techniques may be applied to identify non-uniformity in other regions of a scanned image of a molecular array. The techniques of the present invention may be applied to scanned images of molecular arrays, regardless of the wavelength of light used in an optical scan, energy levels of emitted

25 radiation detected, or other type of signal detection employed to generate the scanned image. Of course, each different type of scanning device, molecular array, type of signal detected, and other variations will need a corresponding variance model for calculating useful variance thresholds.

The foregoing description, for purposes of explanation, used specific

30 nomenclature to provide a thorough understanding of the invention. However, it will be apparent to one skilled in the art that the specific details are not required in order

to practice the invention. The foregoing descriptions of specific embodiments of the present invention are presented for purpose of illustration and description. They are not intended to be exhaustive or to limit the invention to the precise forms disclosed. Obviously many modifications and variations are possible in view of the above

5   teachings. The embodiments are shown and described in order to best explain the principles of the invention and its practical applications, to thereby enable others skilled in the art to best utilize the invention and various embodiments with various modifications as are suited to the particular use contemplated. It is intended that the scope of the invention be defined by the following claims and their equivalents:

10